FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models

Xiaochen Wang[1*], **Jiaqi Wang**[1*], **Houping Xiao**[2], **Jinghui Chen**[1], **Fenglong Ma**[1†]
[1]Pennsylvania State University, [2]Georgia State University
[1]{xcwang, jqwang,jzc5917,fenglong}@psu.edu, [2]hxiao@gsu.edu

Abstract

Foundation models have demonstrated remarkable capabilities in handling diverse modalities and tasks, outperforming conventional artificial intelligence (AI) approaches that are highly task-specific and modality-reliant. In the medical domain, however, the development of comprehensive foundation models is constrained by limited access to diverse modalities and stringent privacy regulations. To address these constraints, this study introduces a novel knowledge injection approach, FEDKIM, designed to scale the medical foundation model within a federated learning framework. FEDKIM leverages lightweight local models to extract healthcare knowledge from private data and integrates this knowledge into a centralized foundation model using a designed adaptive **M**ultitask **M**ultimodal **M**ixture **O**f **E**xperts (M^3OE) module. This method not only preserves privacy but also enhances the model's ability to handle complex medical tasks involving multiple modalities. Our extensive experiments across **twelve** tasks in **seven** modalities demonstrate the effectiveness of FEDKIM in various settings, highlighting its potential to scale medical foundation models without direct access to sensitive data. Source codes are available at https://github.com/XiaochenWang-PSU/FedKIM.

1 Introduction

Similar to large language models (Zhao et al., 2023) and foundation models (Zhou et al., 2023a), medical foundation models (Thirunavukarasu et al., 2023; Moor et al., 2023) have achieved superior performance of handling diverse modalities and tasks within the medical domain. These models have the potential to revolutionize medical diagnostics and treatment by leveraging data-driven insights from large volumes of multimodal healthcare data.

Table 1: Summary of medical foundation models.

Medical Foundation Model	Modalities	Tasks
MMedLM2 (Qiu et al., 2024)	Text	Question-answering
LLava-Med(Liu et al., 2023a)	Text, Image	Visual Question-answering
Med-Flamingo(Yang et al., 2023)	Text, Image	Visual Question-answering
PMC_LLAMA(Lee et al., 2023)	Text	Question-answering
BiomedGPT(Gu et al., 2021)	Text, Image	Visual Question-answering
BioMedLM(Lewis et al., 2020)	Text	Question-answering
GatorTron(Hao et al., 2020)	Text	Clinical concept extraction Medical relation extraction Semantic textual similarity Natural language inference Question-answering
Med-PaLM(Singhal et al., 2022)	Text	Question-answering
ChatDoctor(Li et al., 2023)	Text	Question-answering

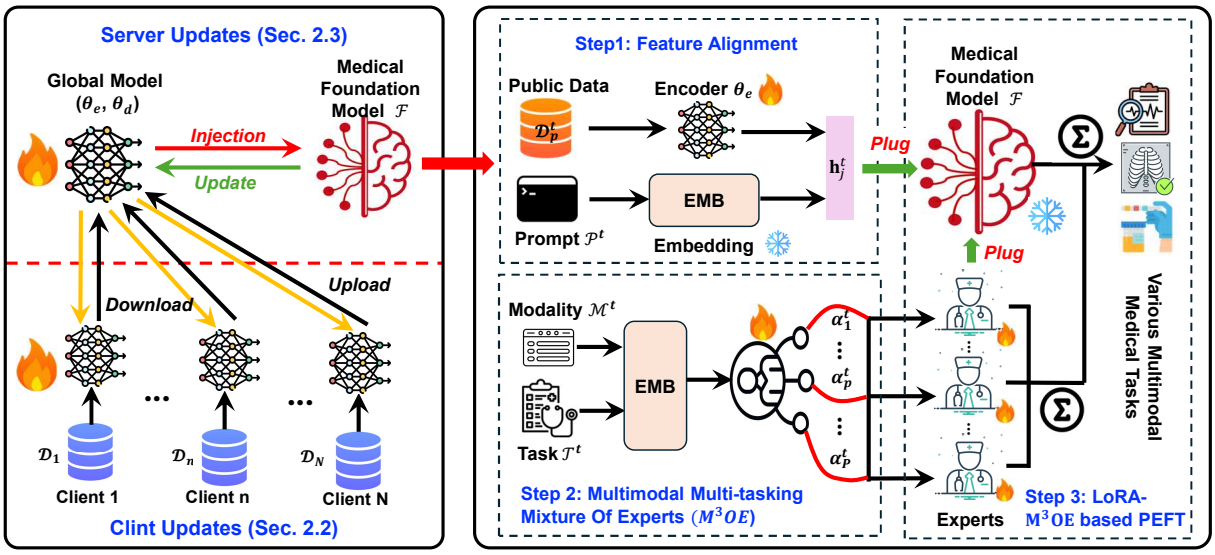
Due to the sensitive nature of medical data and the complexity of medical tasks, most existing medical foundation models usually rely on particular public medical datasets. This nature results in limitations of the existing medical foundation models, detailed as follows:

(1) **Unrealistic to conduct large-scale centralized training**. The centralized training of medical foundation models presents significant challenges, primarily due to the difficulties in aggregating sensitive healthcare data. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union impose strict privacy restrictions on the use of personal health information. This regulatory environment makes it impractical to collect and store large amounts of healthcare data in a single location, which is typically required for the effective training of high-performing medical foundation models.

(2) **Limited modality and task adaptability**. Current medical foundation models exhibit a high degree of specialization, constraining their effectiveness to a narrow range of downstream tasks within specific modalities, as outlined in Table 1. For instance, MMedLM (Qiu et al., 2024) is tailored for text, while LLava-Med (Liu et al., 2023a) focuses on both image and text modalities. In prac-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8141–8154
November 12-16, 2024



(a) Framework Overview (b) Federated Knowledge Injection into Medical Foundation Model

Figure 1: Illustration of the proposed FEDKIM. (a) Framework overview, where the proposed FEDKIM contains client and server updates. (b) Federated knowledge injection, where FEDKIM first aggregates models uploaded from clients and then injects the aggregated model knowledge into medical foundation model $\mathcal{F}$ with three steps. "PEFT" in Step 3 denotes parameter-efficient fine-tuning.

tical settings, comprehensive medical decisions often require integrating multiple types of health data across various tasks. Yet, by being task or modality-specific, existing models fail to recognize and leverage the intricate relationships between different healthcare data modalities and tasks.

The first limitation prevents training a medical foundation model from scratch in a centralized manner, while the second one exacerbates the challenge of developing a multimodal, multi-task medical foundation model. To overcome these obstacles, a viable solution is to scale existing medical foundation models and infuse them with medical knowledge. Given that medical data is stored on private clients, the federated learning (FL) paradigm (McMahan et al., 2017; Che et al., 2023; Zhou et al., 2022) offers a promising approach in the medical domain (Wang et al., 2022a), which is a decentralized and collaborative machine learning method where participants do not need to share data directly. Although several recent studies on federated foundation models (Lu et al., 2023; Chen et al., 2024a) have made progress, they primarily focus on enhancing services to local clients using existing foundation models. Importantly, none have specifically tackled the challenge of injecting novel medical knowledge into existing medical foundation models in a federated manner.

To tackle this new challenge, in this paper, we propose a novel approach: Federated Knowledge Injection for Medical foundation models (FEDKIM), as shown in Figure 1. FEDKIM adopts a flexible design, allowing it to incorporate various types of medical modalities to handle a variety of medical tasks. Considering the real-world scenarios, FEDKIM deploys the medical foundation model only on the server side and leverages lightweight local models along with classic federated learning approaches to extract healthcare knowledge from private data.

To effectively inject extracted medical knowledge into the foundation model, FEDKIM uses knowledge-rich parameters from the modality-specific encoders updated from the local end. To be specific, FEDKIM integrates this knowledge using parameter-efficient fine-tuning technique with a novel multitask multimodal mixture of expert module, namely M^3OE. M^3OE adaptively selects appropriate expert systems for handling specific tasks in given modalities, enabling FEDKIM to deal with tasks in complex medical contexts.

Our experiments across 12 healthcare tasks with 7 modalities demonstrate the effectiveness of FEDKIM, providing a solid foundation for future exploratory research on the medical knowledge injection problem.

2 The proposed FEDKIM Framework

In this section, we first introduce the setup of the medical knowledge injection task (Section 2.1).

Next, we describe the proposed method, FEDKIM. As depicted in Figure 1, FEDKIM consists of two main components: knowledge extractors (Section 2.2), which are deployed on local clients, and a knowledge injector (Section 2.3), which is deployed on the server.

2.1 Framework Setups

2.1.1 Client Setups

The goal of this work is to scale and enhance the predictive ability of medical large language models (LLMs) by incorporating medical knowledge from private client data in a federated manner. To achieve this, we employ N clients, each representing a hospital or a medical institute holding private medical data $\mathcal{D}_n$. We assume that the private dataset $\mathcal{D}_n$ contains all medical modalities $\{\mathcal{M}_1, \cdots, \mathcal{M}_M\}$ and can perform all tasks $\{\mathcal{T}_1, \cdots, \mathcal{T}_T\}$. Each client trains a model $f_n = [\text{ENC}_n(); \text{DEC}_n()]$ using the data $\mathcal{D}_n$, where $\text{ENC}_n()$ is the set of multimodal encoders and $\text{DEC}_n()$ is the set of multi-task decoders/predictors. Thus, the model parameters $\boldsymbol{\theta}_n$ of f_n can be divided into $\boldsymbol{\theta}_n^{enc}$ for the encoder and $\boldsymbol{\theta}_n^{dec}$ for the decoder, which will be further uploaded to the server.

2.1.2 Server Setups

We deploy a generative medical foundation model on the server, denoted as $\mathcal{F}$. We aim to inject medical knowledge represented by $\{\boldsymbol{\theta}_1^{enc}, \cdots, \boldsymbol{\theta}_N^{enc}\}$ into $\mathcal{F}$ and simultaneously update $\{\boldsymbol{\theta}_1^{enc}, \cdots, \boldsymbol{\theta}_N^{enc}\}$ by absorbing new knowledge from $\mathcal{F}$. These updated encoders and the aggregated decoders will then be distributed to the corresponding clients for learning in the next communication round. To facilitate the updates of client parameters, we place a small amount of public data on the server, denoted as $\mathcal{D}_p$.

2.2 Client Updates – Knowledge Extraction from Private Clients

This framework allows each client to handle T tasks simultaneously. Although these tasks have different training data, the modalities are partially shared, which motivates us to design a simple client model with M modality-specific encoders and T task-specific decoders. Details of the encoders are listed in Appendix D. We then use the following loss to train each client model:

$$\min_{\boldsymbol{\theta}_n} \mathcal{L}_n := \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{D}_n^t|}\sum_{(\mathbf{x}_i^t,\mathbf{y}_i^t)\in\mathcal{D}_n^t}\ell^t(f_n(\mathbf{x}_i^t;\boldsymbol{\theta}_n),\mathbf{y}_i^t), \tag{1}$$

$$f_n(\mathbf{x}_i;\boldsymbol{\theta}_n) = \text{DEC}_{n,t}(\text{ENC}_{n,m}(\mathbf{x}_i;\boldsymbol{\theta}_{n,m}^{enc});\boldsymbol{\theta}_{n,t}^{dec}), \tag{2}$$

where $\mathcal{D}_n^t$ is the task-specific dataset, $\mathbf{x}_i^t$ and $\mathbf{y}_i^t$ are the data features and the corresponding ground truths, and ℓ^t is the loss function for a specific task, such as cross-entropy. $\text{ENC}_{n,m} \subseteq \text{ENC}_n$ is the encoder for modality $\mathcal{M}_m$ with parameters $\boldsymbol{\theta}_{n,m}^{enc}$. $\text{DEC}_{n,t} \subseteq \text{DEC}_n$ is the decoder for the t-th task with parameters $\boldsymbol{\theta}_{n,t}^{dec}$. The number of modality-level encoders in $\text{ENC}_{n,m}$ is determined by the input data, while amount of tasks determines the number of task-oriented decoders. After the local training, we will upload the encoder and decoder parameters $\boldsymbol{\theta}_n^{enc}$ and $\boldsymbol{\theta}_n^{dec}$ to the server.

2.3 Server Updates – Knowledge Injection into Medical LLM

2.3.1 Knowledge Aggregation

We assume that the predictive ability of $\mathcal{F}$ is better than the uploaded decoders $\{\boldsymbol{\theta}_1^{dec}, \cdots, \boldsymbol{\theta}_N^{dec}\}$, and useful knowledge is primarily contained in the encoders $\{\boldsymbol{\theta}_1^{enc}, \cdots, \boldsymbol{\theta}_N^{enc}\}$. Thus, on the server side, we aim to inject medical knowledge $\{\boldsymbol{\theta}_1^{enc}, \cdots, \boldsymbol{\theta}_N^{enc}\}$ into the LLM $\mathcal{F}$ with the help of public data $\mathcal{D}_p$. Before the injection, we first aggregate knowledge uploaded from each client in traditional federated learning manners such as FedAvg (McMahan et al., 2017) or FedProx (Li et al., 2020), i.e.,

$$\begin{aligned}\boldsymbol{\theta}_e &= f_{FL}([\boldsymbol{\theta}_e^1, \cdots, \boldsymbol{\theta}_e^M]),\\ \boldsymbol{\theta}_d &= f_{FL}([\boldsymbol{\theta}_d^1, \cdots, \boldsymbol{\theta}_d^M]),\end{aligned} \tag{3}$$

where f_{FL} can be flexibly replaced with any federated learning methods, such as *personalized FL* methods (Jiang et al., 2019; T Dinh et al., 2020), *differential privacy-based FL* methods (Hu et al., 2020; El Ouadrhiri and Abdelhadi, 2022), or *adaptive FL* methods (Reddi et al., 2020; Wang et al., 2022c,b).

2.3.2 Knowledge Injection

Effectively injecting medical knowledge $\boldsymbol{\theta}_e$ is challenging since the LLM $\mathcal{F}$ cannot directly use these diverse modality-specific encoders. To solve this challenge, we leverage a straightforward yet effective feature alignment strategy that follows the

training of LLaVA (Liu et al., 2024) by concatenating the modality embeddings with the task prompt. Subsequently, we embed our original **M**ultimodal **M**ulti-tasking **M**ixture **O**f **E**xperts (M^{3}OE) into the medical foundation model. M^{3}OE allows the medical foundation model $\mathcal{F}$ to adaptively select specific expert system given different combination of tasks and modalities. Next, we detail the process of knowledge injection.

Step 1: Feature Alignment. For each input data $(\mathbf{x}_j^t, \mathbf{y}_j^t) \in \mathcal{D}_p^t$ from the t-th task, we first obtain its feature representations using the aggregated encoders $\boldsymbol{\theta}_e$, i.e., $\mathbf{e}_j^t = [\mathbf{e}_j^1; \cdots; \mathbf{e}_j^M] = g(\boldsymbol{\theta}_e(\mathbf{x}_j^t))$, where $g(\cdot)$ is the linear mapping function. We also embed the task prompt $\mathcal{P}^t$ using the encoder of $\mathcal{F}$, i.e., $\mathbf{p}^t = \text{EMB}_{\mathcal{F}}(\mathcal{P}^t)$., where $\text{EMB}_{\mathcal{F}}()$ is the text embedding layer of $\mathcal{F}$. Then, the concatenation of the data feature $\mathbf{e}_j^t$ and the task prompt feature $\mathbf{p}^t$ will be used as the input of the encoder of $\mathcal{F}$, denoted as $\mathbf{h}_j^t = [\mathbf{e}_j^t; \mathbf{p}^t]$.

Step 2: Multimodal Multi-tasking Mixture of Experts (M^{3}OE). A naive solution is directly using the aligned feature $\mathbf{h}_j^t$ to generate the output. However, such a naive end-to-end fine-tuning approach not only has weak distinguishability of different tasks but also ignores the generalization ability of FEDKIM to unseen tasks, even though the modalities have been encountered already. To address this issue, we develop a **M**ultimodal **M**ulti-tasking **M**ixture **O**f **E**xperts (M^{3}OE) module to allow FEDKIM to distinguish tasks dynamically.

M^{3}OE takes both the task description $\mathcal{T}^t$ and the modality descriptions $\mathcal{M}^t$ associated with Task $\mathcal{T}^t$ as inputs to compute the relevance of each expert for the given task and modality, where $\mathcal{M}^t$ is the concatenation of descriptions of all modalities concerning Task $\mathcal{T}^t$. $\mathcal{T}^t$ and $\mathcal{M}^t$ are firstly encoded by the embedding layer of the foundation model $\mathcal{F}$, and subsequently processed to output weights for expert selection as follows:

$$\begin{aligned}\boldsymbol{\alpha}^t &= \text{softmax}\left(\text{MLP}\left(\text{Pooling}\left(\boldsymbol{\beta}^t\right)\right)\right), \\ \boldsymbol{\beta}^t &= \frac{(\mathbf{W}_q\text{EMB}_{\mathcal{F}}(\mathcal{M}^t))(\mathbf{W}_k\text{EMB}_{\mathcal{F}}(\mathcal{T}^t))^\top}{\sqrt{d_k}}\mathbf{W}_v\text{EMB}_{\mathcal{F}}(\mathcal{T}^t)\end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}^t \in \mathbb{R}^P$ and P is the number of experts. $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ denote the attention matrices, and d_k is the dimension size.

The proposed M^{3}OE effectively integrates the injected knowledge managed by two separate routers, resulting in a more streamlined and contextually aware computation of weights. The output, $\boldsymbol{\alpha}^t$, represents the attention-weighted selection of experts optimized for both the modality and the specific task. This approach provides the flexibility needed to handle complex medical scenarios by selecting the appropriate experts based on the context.

Step 3: LoRA-M^{3}OE based Parameter-Efficient Fine-tuning. Finally, we generate the representation of each layer in $\mathcal{F}$ for the forward pass based on LoRA (Hu et al., 2022) and the learned M^{3}OE weight using Eq. (4) as follows:

$$\mathbf{c}_j^t = \mathbf{W}_{\mathcal{F}}\mathbf{h}_j^t + \sum_{p=1}^{P} \alpha_p^t(\mathbf{B}_p\mathbf{A}_p\mathbf{h}_j^t), \tag{5}$$

where $\mathbf{W}_{\mathcal{F}}$ denotes the frozen parameters of $\mathcal{F}$, $\mathbf{B}_p\mathbf{A}_p$ denotes the lower-rank adaptation module serving as the p-th expert system. We will fine-tune the proposed FEDKIM using the final output from $\mathcal{F}$ and the ground truth $\mathbf{y}_j^t$. The design balances efficacy and efficiency during knowledge injection, allowing FEDKIM to decently handle the complex nature of medical applications.

During the training, modality-specific encoders $\boldsymbol{\theta}_e$ gradually align with the medical LLM $\mathcal{F}$ that contains abundant knowledge acquired through pre-training. The alignment indicates prior knowledge in $\mathcal{F}$ is also extracted during injection, in the form of adjusted parameters restored in encoders. To benefit local models and boost knowledge injection in the next round, FEDKIM passes the updated encoders $\boldsymbol{\theta}_e$ and the aggregated decoders $\boldsymbol{\theta}_d$ back to local ends and performs the knowledge-driven iterative training until convergence.

3 Experiment Setup

3.1 Task Introduction

In this study, we have training tasks and validation tasks across different datasets and data modalities. To provide a clear illustration, we present them in Table 2.

Training Task. To examine the utility of the proposed FEDKIM, we leverage **four** classification tasks across **six** modalities to federatedly inject medical knowledge into the selected foundation model through multi-task training. Details regarding these tasks are available in Appendix A. As emphasized in Section 2, we perform training on this suite of tasks in a multi-task pattern.

Validation Task. Typical medical foundation models, such as MMedLM2 (Qiu et al., 2024), often struggle with handling unseen tasks involving

novel modalities. To evaluate the extent to which knowledge injection enables the medical foundation model to tackle unseen tasks, we compile **five** classification tasks (ECD, SP, PED, AD, and EBD) and **three** generation tasks (MR, SNC, and MS). Details on these tasks are provided in Appendix B.

3.2 Data Partition

For each training task, we divide the data into four parts in a ratio of 7:1:1:1. Specifically, 70% of the data, $\mathcal{D}_n$, is private data evenly distributed to N clients for training local models. Another 10% of the data is public data, $\mathcal{D}_p$, placed on the server for tuning the foundation model. An additional 10% of the data is development data, $\mathcal{D}_d$, kept on the server as a validation set. The remaining 10% of the data, $\mathcal{D}_t$, is used as testing data for these tasks. More details regarding the data distribution can be found in Appendix C.

3.3 Baselines

Since the task of medical knowledge injection is novel and unexplored, there are no existing baselines. Therefore, we establish our own baselines, detailed as follows:

FedPlug. FedPlug acquires modality-specific encoders through the federated learning process described in Section 2.2 . These encoders are then integrated into the foundation model for fine-tuning. By aligning multimodal medical input with the semantic space of the foundation model, FedPlug enables the model to handle multiple modalities. Throughout this process, only the aggregated encoders are trainable.

FedPlug$_L$. Building on the FedPlug framework, FedPlug$_L$ incorporates the Low-Rank Adaptation (LoRA) technique (Hu et al., 2022) to better integrate multimodal features into the semantic space of the large language model (LLM), thereby optimizing the federated learning process. In addition to the trainable encoders in FedPlug, each layer of the LLM is equipped with a tunable LoRA module.

3.4 FL Backbone Approaches

We implement our FEDKIM based upon the following backbone approaches:

FedAVG (McMahan et al., 2017) is a conventional federated learning method, producing a global model by aggregating distributed models $[\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^M]$ as follows:

$$f_{avg}([\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^M]) = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\theta}^n.$$

FedProx (Li et al., 2020) aims to extend FedAvg by regularizing each local loss function with an L_2 term as follows:

$$\min_{\boldsymbol{\theta}^n} \mathcal{J}_n(\boldsymbol{\theta}^n; \boldsymbol{\theta}^*) = \mathcal{L}_n(\boldsymbol{\theta}^n) + \frac{\lambda}{2}||\boldsymbol{\theta}^n - \boldsymbol{\theta}^*||^2, \quad (6)$$

where $\boldsymbol{\theta}^*$ is the global model, $\mathcal{L}_n(\cdot)$ is the corresponding loss function, and λ is the hyperparameter for weighting.

MMedLM-2[1] (Qiu et al., 2024) is an advanced unimodal Large Language Model. Benefiting from multilingual pre-training, MMedLM-2 achieves the state-of-the-art performance in multiple question answering tasks, thus selected as the backbone of our foundation model deployed on the server.

3.5 Implementation Details

All experiments were conducted in an Ubuntu 20.04 environment using two NVIDIA A100 GPUs. We utilized MMedLM-2, the aforementioned state-of-the-art pre-trained medical language model, as the target of medical knowledge injection. The learning rate was set to 5×10^{-4} for the foundation model and 1×10^{-4} for the local models. λ for FedProx was set to 1×10^{-4}. Cross-entropy loss was used for training the local models, while the foundation model was optimized using general autoregressive loss. The number of clients N was set to 5, and the number of experts P was set to 12 for FEDKIM. To ensure a fair comparison, we set the number of communication rounds to 10 for all methods involved in the comparison.

4 Performance Evaluation

We examine our proposed FEDKIM from the zero-shot evaluation (subsection 4.1) and fine-tuning evaluation (subsection 4.2) perspectives.

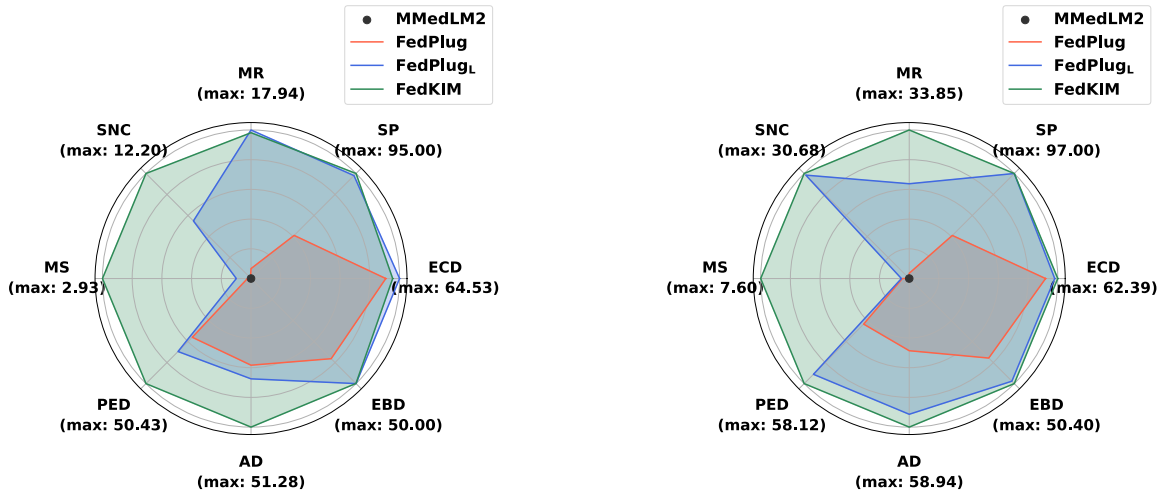
4.1 Zero-shot Evaluation

In the zero-shot evaluation, there is no overlap between the training tasks and evaluation tasks, which targets at examining the zero-shot capability of the medical foundation models enabled by FEDKIM. The experiment results on unseen tasks are shown in Figure 2, with FedAvg (Figure 2a) and FedProx

[1] https://huggingface.co/Henrychur/MMedLM2

Table 2: Tasks and modalities in this study.

Task Type	Task	Modality						
		Image	Signal	Vital signs	Lab events	Input	Output	Text
Training	COVID-19 Detection (**CD**)	✓	✗	✗	✗	✗	✗	✗
	Lung Opacity Detection (**LOD**)	✓	✗	✗	✗	✗	✗	✗
	ECG Abnormal Detection (**EAD**)	✗	✓	✗	✗	✗	✗	✗
	Mortality Prediction (**MP**)	✗	✗	✓	✓	✓	✓	✗
Validation	Enlarged Cardiomediastinum Detection (**ECD**)	✓	✗	✗	✗	✗	✗	✗
	Pleural Effusion Detection (**PED**)	✓	✗	✗	✗	✗	✗	✗
	Atelectasis Detection (**AD**)	✓	✗	✗	✗	✗	✗	✗
	Ectopic Beats Detection (**EBD**)	✗	✓	✗	✗	✗	✗	✗
	Sepsis Prediction (**SP**)	✗	✗	✓	✓	✓	✓	✗
	MedVQA-RAD (**MR**)	✓	✗	✗	✗	✗	✗	✓
	MedVQA-Slake (**MS**)	✓	✗	✗	✗	✗	✗	✓
	Signal Noise Clarification (**SNC**)	✗	✓	✗	✗	✗	✗	✓



(a) FedAvg-based Knowledge Injection Performance.

(b) FedProx-based Knowledge Injection Performance.

Figure 2: Performance comparison between FEDKIM and baselines on the zero-shot evaluation.

(Figure 2b) as the backbone federated approaches. We use black •, orange, blue, and green curves to denote MMedLM2, FedPlug, FedPlug$_L$, and FEDKIM, respectively. Accuracy for classification tasks and BLEU for generation tasks are used for visualization. Based on the experiment results, we provide the observations and discussion below:

(1) The original foundation model MMedLM2 fails to do the zero-shot evaluation on the unseen tasks in the training process. This is due to its extremely limited multimodal capabilities.

(2) FedPlug, which only incorporates the federated encoder, performs the worst across all tasks, regardless of the type. This observation underscores the necessity of effectively utilizing public data to align the medical foundation model with external knowledge. Without proper integration, external knowledge—although derived through federated approaches on vast amounts of private data—cannot be directly assimilated into the medical foundation model.

(3) Even though FedPlug$_L$ approaches FEDKIM's performance on several tasks, it still falls short, particularly in generation tasks like MedVQA. This indicates that the knowledge injected through FedPlug+LoRA does not fully generalize to unseen tasks, as training was exclusively performed on classification tasks. In contrast, FEDKIM, despite also being trained on classification tasks, achieves better performance on these tasks and maintains superior capability in handling unseen classification tasks. Comparing our FEDKIM with FedPlug$_L$, FEDKIM shows the superior performance on all the tasks, especially on the tasks of SNC ($\uparrow$ 82.36% with FedAvg), PED ($\uparrow$ 43.92% with FedAvg), and AD ($\uparrow$ 48.12%). On the other tasks, such as MR, EBD, and ECD, these approaches reach closed performance. This success is attributed to the M^{3}OE module, which enables FEDKIM to adaptively select appropriate experts to jointly handle novel tasks based on the context. Furthermore, our proposed FEDKIM works well with the federated backbones of FedAvg and FedProx. It also generally maintains the advantages of

a more advanced federated learning method (FedProx) over the vanilla approach. Comparing Figure 2a and Figure 2b, the performance with FedProx generally outperforms the one with FedAvg on different tasks, such as PED $\uparrow 15.25\%$.

These observations further show the adaptability of FEDKIM to enable medical foundation models to have zero-shot capability across different tasks and federated learning frameworks.

4.2 Fine-tuning Evaluation

While injecting medical knowledge into foundation models demonstrates the potential for handling unseen tasks, it remains uncertain whether the enhanced foundation model can also perform well on previously encountered tasks. To address this, we conducted a fine-tuning evaluation, with the training process detailed in Section 3.1 considered as fine-tuning for these tasks. The test sets for these tasks were used for evaluation, and the fine-tuning results are presented in Table 3. For a comprehensive evaluation, we utilize accuracy, precision, recall, and F1 score as metrics for these tasks.

Compared to the experiments on unseen tasks, it is evident that the knowledge-injected medical foundation model performs significantly better on familiar tasks. This showcases the explicit utilization of knowledge acquired through federated training. Similar to the zero-shot evaluation, approaches combined with FedProx consistently outperform those with FedAvg, underscoring the importance of effective knowledge extraction during the injection process.

Furthermore, FEDKIM consistently outperforms the two baselines, FedPlug and FedPlug$_L$. This competitive performance validates the design and effectiveness of the M^{3}OE module.

4.3 Ablation Study

We conduct ablation studies on the COVID-19 detection and enlarged cardiomediastinum detection tasks to assess the impact of each module within our proposed FEDKIMin both fine-tuning and zero-shot settings. Retaining all other modules as in the main experiments, we explore the following variant settings: (1) FEDKIM^{pub}: Instead of utilizing knowledge from private datasets $\mathcal{D}_n$, this configuration solely leverages public dataset $\mathcal{D}_p$ for centralized training. Consequently, the federated training module discussed in Section 2.2 is excluded, with the encoder $\boldsymbol{\theta}_e$ updated exclusively through public training as detailed in Section 2.3. (2) FEDKIM$^{\mathcal{T}}$: This variant omits the task description module that guides the expert selection process, testing the importance of task-specific information in routing the mixture of experts. (3) FEDKIM$^{\mathcal{M}}$: Similarly, we remove the modality description module to examine its influence on expert selection.

The results of the ablation studies are presented in Table 4 and Table 5. They indicate that each component significantly enhances FEDKIM's performance. Specifically, a substantial decline in performance with FEDKIM^{pub} highlights the crucial role of knowledge injected from local clients through federated learning. This locally enriched encoder allows the medical foundation model to better adjust to unseen modalities, thereby enhancing its effectiveness compared to models trained without this knowledge. Moreover, the absence of task or modality descriptions diminishes FEDKIM's ability to manage specific tasks through multi-task training, validating the design of the M^{3}OE module. This module equips FEDKIM to effectively navigate complex healthcare scenarios that involve diverse tasks and modalities. In summary, the synergistic integration of local knowledge, along with the task and modality description modules, crucially bolsters the performance of our proposed FEDKIM.

5 Related Work

Medical Foundation Models. Foundation models, known for their vast parameters and training datasets, have demonstrated impressive capabilities across domains (Touvron et al., 2023; Zhou et al., 2023a; Yang et al., 2024; Li et al., 2024a; Abbasian et al., 2024), and are becoming increasingly prevalent in healthcare. Thirunavukarasu et al. (Thirunavukarasu et al., 2023) highlight the potential of large language models (LLMs) in clinical settings. Moor et al. (Moor et al., 2023) propose a generalist medical AI for diverse tasks using multimodal data. Specialized medical foundation models have been developed for disease detection (Zhou et al., 2023b), cancer biomarker identification (Pai et al., 2024), echocardiogram interpretation (Christensen et al., 2024), image segmentation (Zhang et al., 2023a), and precision oncology (Truhn et al., 2024). Despite these advancements, this area remains relatively unexplored compared to the general domain (Wang et al., 2024a), primarily due to the complexity inherent in healthcare data (Wang et al., 2023, 2024b).

Table 3: Fine-tuning evaluation for training tasks. ✗ denotes incapacity.

Task	Method	LLM	FedAvg			FedProx		
	Metric	MMedLM-2	FedPlug	FedPlug$_L$	FEDKIM	FedPlug	FedPlug$_L$	FEDKIM
Covid-19 Detection	Accuracy	✗	98.34	94.21	98.48	86.11	95.73	98.98
	Precision	✗	96.07	99.27	96.61	65.09	99.32	98.28
	Recall	✗	97.44	77.78	97.44	97.72	83.76	97.72
	F1	✗	96.75	87.22	97.02	78.13	90.88	98.00
Lung Opacity Detection	Accuracy	✗	95.48	85.45	94.99	93.13	78.64	95.10
	Precision	✗	98.22	99.85	98.32	93.74	72.20	97.15
	Recall	✗	92.95	74.03	91.90	92.95	95.58	93.27
	F1	✗	95.51	83.69	95.00	93.35	82.26	95.17
ECG Abnormal Detection	Accuracy	✗	43.15	42.28	44.75	45.25	50.94	58.46
	Precision	✗	56.97	82.61	61.11	60.85	58.01	58.46
	Recall	✗	11.22	1.49	13.80	17.80	58.20	100.00
	F1	✗	18.74	2.93	22.52	27.55	58.10	73.78
Mortality Prediction	Accuracy	✗	84.11	53.63	90.01	82.41	91.42	89.99
	Precision	✗	16.35	10.96	35.88	13.87	47.57	36.97
	Recall	✗	21.43	63.04	23.29	16.64	15.22	27.33
	F1	✗	18.55	18.67	28.24	15.13	23.06	31.43

Table 4: Ablation study results in the fine-tuning setting.

Setting	Accuracy	Precision	Recall	F1
FEDKIM^{pub}	94.07	90.63	85.47	87.98
FEDKIM$^{\mathcal{T}}$	98.34	96.28	97.12	96.70
FEDKIM$^{\mathcal{M}}$	98.41	97.13	96.58	96.86
FEDKIM	98.48	96.61	97.44	97.02

Table 5: Ablation study results in the zero-shot setting.

Setting	Accuracy	Precision	Recall	F1
FEDKIM^{pub}	59.82	54.43	84.40	66.19
FEDKIM$^{\mathcal{T}}$	61.11	55.80	79.82	65.66
FEDKIM$^{\mathcal{M}}$	60.26	66.00	30.28	41.51
FEDKIM	61.54	55.13	93.58	69.39

Federated Fine-tuning with Foundation Models. Fine-tuning foundation models (FMs) with task-specific data is essential for improved performance in specialized tasks. Federated Learning (FL) supports this by utilizing locally stored data and distributed computational resources. Research in this field includes full tuning (Deng et al., 2023; Fan et al., 2023), partial tuning (Peng et al., 2024; Marchisio et al., 2022; Khalid et al., 2023), and parameter-efficient fine-tuning (PEFT) (Lu et al., 2023; Zhang et al., 2023b). Notably, (Lu et al., 2023) involves clients hosting FMs and exchanging adapters with the server, which aggregates and redistributes them. Similarly, FedPETuning (Zhang et al., 2023b) shares parts of client models for pre-trained language models in FL. Unlike these studies, which require clients to have FMs, our approach positions the medical FM on the server, facilitating collaborative enhancement of medical FM models without accessing local data.

Parameter-efficient Fine-tuning on Foundation Model Full-parameter Fine-Tuning of foundation models, while promising in terms of performance enhancement, requires extremely extensive computational resources. Consequently, researchers have investigated Parameter-efficient Fine-tuning (PEFT) techniques. PEFT methods aim to adapt pre-trained models to specific tasks using a minimal number of additional parameters. Low-Rank Adaptation (LoRA) (Hu et al., 2022), a widely recognized PEFT method, reduces the number of trainable parameters by factorizing weight matrices into low-rank representations, achieving significant parameter efficiency. Additionally, previous studies have utilized modular approaches, such as adapters (Gao et al., 2023) and the Perceiver Resampler (Alayrac et al., 2022), to adapt new modalities to foundation models.

Researchers have explored combining the Mixture of Experts (Jacobs et al., 1991) concept with Low-Rank Adaptation (LoRA) for Parameter-efficient Fine-tuning (PEFT) (Li et al., 2024b; Wu et al., 2023). To guide the selection of experts in complex scenarios, they have leveraged modality information (Luo et al., 2024; Li et al., 2024c), instructions (Chen et al., 2023, 2024b; Wu et al., 2023; Li et al., 2024b), or pre-defined task IDs (Liu et al., 2023b). However, these MOE methodologies do not specifically address the complex, modality-diverse scenarios found in the healthcare domain.

6 Conclusion

This work introduces the concept of knowledge injection into medical foundation models, emphasizing its critical role and potential in the development of comprehensive medical models. We propose a novel approach, FEDKIM, designed to extract and inject healthcare knowledge into foundation

models, thereby enhancing their ability to handle multiple tasks and modalities. FEDKIM leverages flexible federated learning techniques to extract knowledge from distributed medical data. The extracted knowledge is then injected into the foundation model using our proposed adaptive M^3OE module. Our exhaustive experimental results on 12 tasks and 7 modalities demonstrate the effectiveness of FEDKIM in diverse settings, showcasing its excellent capability in handling either encountered or unseen healthcare tasks. This study validates the potential of injecting knowledge into foundation models using federated learning, providing a crucial solution for developing a healthcare foundation model without accessing sensitive data.

7 Limitations

This work explores the problem of medical knowledge injection within the PEFT framework. Due to current computational limitations, we have not yet combined Full-parameter Fine-Tuning with our proposed FEDKIM. Additionally, our study utilizes MMedLM2, which has 7 billion parameters, but injecting knowledge into larger foundation models is restricted by available computational resources. In future research, we plan to investigate the integration of knowledge injection with Full-parameter Fine-Tuning. We also aim to evaluate the efficacy of our approach on larger medical foundation models to further validate its scalability and potential.

Acknowledgements

This work is partially supported by the National Science Foundation under Grant No. 2238275 and 2348541 and the National Institutes of Health under Grant No. R01AG077016.

References

Rsna pneumonia detection challenge (2018). https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018. Accessed: 2024-06-05.

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, et al. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine*, 7(1):82.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. 2023. Multimodal federated learning: A survey. *Sensors*, 23(15):6986.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. 2024a. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11285–11293.

Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024b. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. 2023. Octavius: Mitigating task interference in mllms via moe. *arXiv preprint arXiv:2311.02684*.

Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. 2024. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8.

Yongheng Deng, Ziqing Qiao, Ju Ren, Yang Liu, and Yaoxue Zhang. 2023. Mutual enhancement of large and small language models with cross-silo knowledge transfer. *arXiv preprint arXiv:2312.05842*.

Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380.

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. 2023. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Yu Gu, Robert Tinn, Hao Cheng, Youzheng Ben, Zhuozhao Liu, Jingqi Zhou, Michael Wang, Shizhuo Wang, Hongfang Zhou, and Yanshan Shen. 2021. Biomedgpt: A large-scale biomedical generative pretrained transformer for biomedical text mining. *arXiv preprint arXiv:2104.07497*.

Tianxiao Hao, Junyi Jessy Wang, Chengyu Liu, Haoran Zhang, Junjie Hu, Haomin Deng, Longxiang Ding, Yitao Si, Yue Gong, Xinyu Han, et al. 2020. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2010.16114*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. 2020. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. 2019. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Umar Khalid, Hasan Iqbal, Saeed Vahidian, Jing Hua, and Chen Chen. 2023. Cefhri: A communication efficient federated learning framework for recognizing industrial human-robot interaction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10141–10148. IEEE.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jinhyuk Lee, Wonjin Yoon, Donghyeon Kim, Youngjoong Jo, and Jaewoo Kang. 2023. Pmc-llama: Large language models for biomedical literature. *arXiv preprint arXiv:2304.07930*.

Michael Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Biomedlm: Large-scale biomedical language model. *arXiv preprint arXiv:2004.03982*.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024a. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024b. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Fedprox: A federated learning optimization algorithm. In *Proceedings of the 2nd MLSys Conference*.

Xiang Li, Yifan Wang, Xiaoman Zhang, Haoran Lin, Junyi Liu, Xiaoxuan Jiang, Weixiong Lin, and Yang Zhang. 2023. Chatdoctor: A doctor-patient dialogue system. *arXiv preprint arXiv:2301.01285*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2024c. Uni-moe: Scaling unified multimodal llms with mixture of experts. *arXiv preprint arXiv:2405.11273*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023b. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *arXiv preprint arXiv:2310.18339*.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. 2023. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2024. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2022. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. *arXiv preprint arXiv:2212.10503*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. 2024. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36.

Suraj Pai, Dennis Bontempi, Ibrahim Hadzic, Vasco Prudente, Mateo Sokač, Tafadzwa L Chaunzwa, Simon Bernatz, Ahmed Hosny, Raymond H Mak, Nicolai J Birkbak, et al. 2024. Foundation model for cancer imaging biomarkers. *Nature machine intelligence*, pages 1–14.

Zhaopeng Peng, Xiaoliang Fan, Yufan Chen, Zheng Wang, Shirui Pan, Chenglu Wen, Ruisheng Zhang, and Cheng Wang. 2024. Fedpft: Federated proxy fine-tuning of foundation models. *arXiv preprint arXiv:2404.11536*.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. 2021. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319.

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. 2020. Adaptive federated optimization. In *International Conference on Learning Representations*.

Karan Singhal, Shekoofeh Azizi, Tong N Tu, Soroush S Mahdavi, Jason T Wei, Hyung Won Chung, Nathan Scales, Aneesh Tanwani, Heather Cole-Lewis, Stephen R Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. 2020. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Daniel Truhn, Jan-Niklas Eckardt, Dyke Ferber, and Jakob Nikolas Kather. 2024. Large language models and multimodal foundation models for precision oncology. *NPJ Precision Oncology*, 8(1):72.

Robin van de Water, Hendrik Nils Aurel Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. 2024. Yet another icu benchmark: A flexible multi-center framework for clinical ml. In *The Twelfth International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15.

Jiaqi Wang, Junyu Luo, Muchao Ye, Xiaochen Wang, Yuan Zhong, Aofei Chang, Guanjie Huang, Ziyi Yin, Cao Xiao, Jimeng Sun, and Fenglong Ma. 2024a. Recent advances in predictive modeling with electronic health records. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8272–8280.

Jiaqi Wang, Cheng Qian, Suhan Cui, Lucas Glass, and Fenglong Ma. 2022a. Towards federated covid-19 vaccine side effect prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 2839. NIH Public Access.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Yuan Zhong, Xiaokun Zhang, Yaqing Wang, Parminder Bhatia, Cao Xiao, and Fenglong Ma. 2024b. Unity in diversity: Collaborative pre-training across multimodal medical sources. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3644–3656.

Yujia Wang, Lu Lin, and Jinghui Chen. 2022b. Communication-compressed adaptive gradient method for distributed nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 6292–6320. PMLR.

Yujia Wang, Lu Lin, and Jinghui Chen. 2022c. Communication-efficient adaptive federated learning. In *International Conference on Machine Learning*, pages 22802–22838. PMLR.

Xun Wu, Shaohan Huang, and Furu Wei. 2023. Mole: Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*.

Jianlin Yang, Yan Zhang, Li Li, Haifeng Wang, Tianyu Liu, and Hua Wu. 2023. Med-flamingo: a multimodal medical foundation model. *Preprint*, arXiv:2304.08100.

Weikai Yang, Mengchen Liu, Zheng Wang, and Shixia Liu. 2024. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, pages 1–26.

Yizhe Zhang, Tao Zhou, Shuo Wang, Peixian Liang, Yejia Zhang, and Danny Z Chen. 2023a. Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–139. Springer.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023b. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Annual Meeting of the Association of Computational Linguistics 2023*, pages 9963–9977. Association for Computational Linguistics (ACL).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023a. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Yao Zhou, Jun Wu, Haixun Wang, and Jingrui He. 2022. Adversarial robustness through bias variance decomposition: A new perspective for federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 2753–2762. ACM.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. 2023b. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163.

A Details of Training Tasks

COVID-19 Detection (CD) involves identifying COVID-19 symptoms from X-ray images using the COVQU dataset (Rahman et al., 2021) to evaluate the model's ability to interpret medical images.
Lung Opacity Detection (LOD) uses chest X-ray images to classify lung opacity based on data from the RSNA Pneumonia Detection Challenge 2018 (rsn), annotated by medical practitioners.
ECG Abnormal Detection (EAD) is an unimodal binary classification task that determines abnormal patterns in 10-second, 12-lead ECG signals from PTB-XL database (Wagner et al., 2020).
Mortality Prediction (MP) predicts ICU patient survival or death using multimodal dynamic features vital signs, lab tests, input and output, with data sourced from MIMIC-III (Johnson et al., 2016).

B Details of Validation Tasks

Enlarged Cardiomediastinum Detection (ECD) (Irvin et al., 2019) aims to assess the presence of an enlarged cardiomediastinum using medical images from clinical evaluations. This task measures the model's capability to interpret radiographic data effectively.
Sepsis Prediction (SP) aims to forecast the likelihood of sepsis during ICU stays, testing the model's ability to understand various clinical features. These features are identical to those used in the mortality prediction task, extracted from the MIMIC-III database through the preprocessing pipeline (van de Water et al., 2024).
Medical Visual Question Answering on RAD (MR) involves using both visual images and textual questions as inputs to generate answers. This task evaluates the model's ability to align text and image modalities within the medical domain. The VQA-RAD dataset is utilized for this task (Lau et al., 2018).

Table 6: Details about the datasets.

Task Type	Task	Total Samples	Private Clients	Public (Server)	Development	Testing
Training Tasks	Lung Opacity Detection	18,406	12,880	1,849	1,841	1,836
	COVID-19 Detection	13,808	9,665	1,380	1,380	1,383
	ECG Abnormal Detection	21,797	15,259	2,179	2,180	2,179
	Mortality Prediction	38,129	26,690	3,812	3,812	3,813
Validation	Enlarged Cardiomediastinum Detection	234	✗	✗	✗	234
	Pleural Effusion Detection	234	✗	✗	✗	234
	Atelectasis Detection	234	✗	✗	✗	234
	Sepsis Prediction	1,000	✗	✗	✗	1,000
	MedVQA-RAD	1,000	✗	✗	✗	1,000
	MedVQA-Slake	1,000	✗	✗	✗	1,000
	Signal Noise Clarification	1,000	✗	✗	✗	1,000
	Ectopic Beats Detection	2,000	✗	✗	✗	2,000

Table 7: Single task fine-tuning evaluation.

Task	Method	LLM	FedAvg			FedProx		
	Metric	MMedLM-2	FedPlug	FedPlug$_L$	FEDKIM	FedPlug	FedPlug$_L$	FEDKIM
COVID-19 Detection	Accuracy	✗	92.34	99.56	99.57	84.16	95.66	98.91
	Precision	✗	93.59	98.87	99.15	77.27	84.18	100.00
	Recall	✗	74.92	99.43	99.15	53.27	98.68	95.73
	F1	✗	79.15	99.14	99.15	63.07	90.85	97.82
Lung Opacity Detection	Accuracy	✗	89.42	51.69	94.93	91.23	90.90	93.63
	Precision	✗	84.69	51.74	93.60	87.76	88.06	92.55
	Recall	✗	97.16	99.79	96.85	96.52	95.37	95.37
	F1	✗	90.50	68.10	95.19	91.93	91.57	93.94
ECG Abnormal Detection	Accuracy	✗	43.15	48.79	58.46	45.25	50.80	58.00
	Precision	✗	56.97	65.02	58.46	60.85	58.47	58.34
	Recall	✗	11.22	26.82	100.00	17.80	54.67	98.51
	F1	✗	18.74	37.98	73.78	27.55	56.51	73.28
Mortality Prediction	Accuracy	✗	84.11	91.67	64.16	82.41	90.98	57.38
	Precision	✗	16.35	43.24	12.86	13.87	40.68	13.25
	Recall	✗	21.43	14.91	56.21	16.64	14.91	72.98
	F1	✗	18.55	22.18	20.94	15.13	21.82	22.42

Signal Noise Clarification (SNC) is a generative task that focuses on accurately describing noise in ECG signals based on corresponding textual questions. The data is extracted from an existing ECG question-answering dataset (Oh et al., 2024). The signals are recorded in 12 channels and last for 10 seconds, similar to the ECG Abnormal Detection task.

Pleural Effusion Detection (PED) (Irvin et al., 2019) is derived from the CheXpert dataset and involves using X-ray images to identify the presence of pleural effusion, testing the model's ability to interpret radiographic data.

Atelectasis Detection (AD) (Irvin et al., 2019) also uses X-ray images from the CheXpert dataset to detect atelectasis, evaluating the model's capability in analyzing medical images.

Medical Visual Question Answering on Slake (MS) (Liu et al., 2021) utilizes both visual images and textual questions from the SLAKE dataset to generate answers, assessing the model's proficiency in aligning text and image modalities in the medical domain.

Ectopic Beats Detection (EBD) aims to identify ectopic beats in ECG signals sourced from the PTB-XL database (Wagner et al., 2020).

C Dataset Details

For tasks involved in training, we adopt the data partition setup detailed in Section 3.2. For tasks utilized in zero-shot evaluation, we select a subset of corresponding datasets to facilitate the inference efficiency. We cover 1,000 randomly sampled samples for the tasks of Sepsis Prediction, MedVQA-

Slake, MedVQA-RAD, Signal Noise Clarification. For Ectopic Beats Detection, we cover 1,000 positive cases and 1,000 negative cases randomly sampled from the dataset, as the original annotations concerning ectopic beats are highly sparse. For Enlarged Cardiomediastinum Detection, Pleural Effusion Detection and Atelectasis Detection, we leverage an existing validation set, which involves 234 samples. Statistics about the datasets leveraged in this study are available in Table 6.

D Modality Encoding

We list all encoders along with corresponding modalities in Table 8. Note that all these encoders can be flexibly replaced with other qualified encoders under the framework of FEDKIM.

Table 8: Details of modality-specific encoders.

Modality	Encoder
Image	Deit-tiny (Touvron et al., 2021)
Signal	CNN (LeCun et al., 1998)
Vital Sign	Transformer (Vaswani et al., 2017)
Lab Results	Transformer (Vaswani et al., 2017)
Input	Transformer (Vaswani et al., 2017)
Output	Transformer (Vaswani et al., 2017)

E Single-task Fine-tuning Evaluation

In addition to the evaluation discussed in Section 4, another auxiliary topic worth investigating is whether FEDKIM can enhance a model's performance on a single task. This is particularly meaningful for practitioners who aim to address a specific task, given the scarcity and specialization often associated with medical data. Therefore, we designed a single-task fine-tuning evaluation, where FEDKIM is applied to the foundation model for each individual training task. The experimental results are presented in Table 7.

The results verify that FEDKIM, benefiting from a well-designed knowledge injection strategy, outperforms both baselines in most tasks. This exploration demonstrates the applicability of FEDKIM even when tasks for injection are limited, thereby broadening its application scope in complex medical scenarios.